Artificial Fingerprinting for Generative Models: Rooting Deepfake Attribution in Training Data

Year: 2020 | Citations: 263 | Authors: Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, Mario Fritz

Abstract

Photorealistic image generation has reached a new level of quality due to the breakthroughs of generative adversarial networks (GANs). Yet, the dark side of such deepfakes, the malicious use of generated media, raises concerns about visual misinformation. While existing research work on deepfake detection demonstrates high accuracy, it is subject to advances in generation techniques and adversarial iterations on detection countermeasure techniques. Thus, we seek a proactive and sustainable solution on deepfake detection, that is agnostic to the evolution of generative models, by introducing artificial fingerprints into the models.Our approach is simple and effective. We first embed artificial fingerprints into training data, then validate a surprising discovery on the transferability of such fingerprints from training data to generative models, which in turn appears in the generated deepfakes. Experiments show that our fingerprinting solution (1) holds for a variety of cutting-edge generative models, (2) leads to a negligible side effect on generation quality, (3) stays robust against image-level and model-level perturbations, (4) stays hard to be detected by adversaries, and (5) converts deepfake detection and attribution into trivial tasks and outperforms the recent state-of-the-art baselines. Our solution closes the responsibility loop between publishing pre-trained generative model inventions and their possible misuses, which makes it independent of the current arms race.